Written evidence submitted by Haydn Belfield, Amritha Jayanti and Dr Shahar Avin, University of Cambridge's Centre for the Study of Existential Risk

Defence industrial policy: procurement and prosperity

1. Executive Summary

1.1 We are researchers at the University of Cambridge's Centre for the Study of Existential Risk.

1.2 In this response we particularly focus on defence and those in adjacent markets systems that integrate increasingly capable artificial intelligence (AI), especially those based on machine learning (ML). Many systems that the Ministry of Defence (MoD) is likely to procure over the next 5-10 years will integrate AI and ML; these systems are likely to both be strategically important and to introduce new vulnerabilities [1][2][3]. These vulnerabilities are likely to pose significant national security risks over the next few decades, both for the UK and the UK's allies. These systems are the focus of much of our work [4][5][6][7], and where we hope to add our expertise to the Committee's Inquiry.

1.3 From our research and interactions with defence and procurement practitioners, we draw the following conclusions:

- Militaries worldwide are beginning, and will likely continue, to procure systems that integrate increasingly capable AI and ML to deliver greater speed, capability or other purported defence advantages.
- However, if these systems are procured and deployed 'prematurely' - before they are fully technologically ready, derisked, safe and secure - they could introduce several new vulnerabilities, including safety, security and systemic risks.
- The market for these systems is characterised by: leadership by the private sector; dominance at the infrastructure level and in R&D by a handful of multinationals; rapid progress and obsolescence cycles meaning most systems are novel; and a development environment in which safety and security at the level needed for defence are rarely present.
- These market characteristics, especially the novelty of the systems and private sector leadership, have contributed to the potential for a skills gap within the MoD while the ability to understand the risks and system readiness of these systems during procurement may not always be present.
- The narrative of safe and responsible autonomous defence systems focuses on the end-user human operator. This focus on the end user is necessary but not sufficient. Risks need to be mitigated at all stages of a system's life-cycle, especially procurement.

- The MoD's idiosyncratic definition of lethal autonomous weapons systems is holding the UK back from providing global leadership and creating uncertainty for the UK's procurement decisions.

1.4 Combined, this leads to risks and oversights in supply and procurement - specifically the risk that the MoD prematurely will procure and deploy defence systems that integrate AI and ML, including in ways that affect strategic operations, thus introducing both known and unknown vulnerabilities.

1.5 We therefore make the following recommendations to protect against premature and/or unsafe procurement and deployment of ML-based systems:

A. Improve systemic risk assessment in defence procurement.
B. Ensure clear lines of responsibility so that senior officials can be held responsible for errors caused in the procurement chain and are therefore incentivised to reduce them;
C. Acknowledge potential shifts in international standards for autonomous systems, and build flexible procurement standards accordingly.
D. Update the MoD's definition of lethal autonomous weapons - the Integrated Security, Defence and Foreign Policy Review provides an excellent opportunity to bring the UK in line with its allies.

1.6 Given the broad scope of the questions asked by the Committee, we are submitting a thematic response to the Committee's inquiry. Our response is centred around the following four of the Committee's questions:

- What are the **national skills and competencies** needed for a successful UK defence industrial sector? How can the UK ensure, and assure, that these are maintained in the right place at the right time for the right cost?
- Does the **market** for Defence systems, products and services have any **specific characteristics**, which differentiates it from other markets?
- Does the MoD understand the **risks and opportunities in the Defence supply chain**, and the procurement strategies of other buyers in the market?
- Given that major capability acquisition programmes are international by design how does a modern national defence research and industrial policy successfully manage **cross-border long term partnerships** and **align with the industrial approach of allies** and partners? What lessons can be learnt from other defence exporting countries?

2. <u>Background</u>

2.1 Our assessment is that if sensible precautions are not taken, militaries are beginning, and will likely continue, to prematurely procure and deploy AI and ML-based systems, including in ways that affect strategic operations. If these systems are procured and deployed ‘prematurely’ - before they are fully technologically ready, derisked, safe and secure - they could introduce several significant new vulnerabilities, including safety, security and systemic risks

2.2 Examples of current defence and defence-adjacent systems that integrate increasingly capable AI and ML include:

- The 50+ systems demonstrated at Unmanned Warrior (2016), and almost 70 systems demonstrated at Exercise Autonomous Warrior (2018).
- Several systems funded by the Autonomy in a Dynamic World competition run by the Defence and Security Accelerator, or under development at NavyX.

2.3 Systems of this nature are a priority for the MoD [8][9]. Over the next 5 to 10 years, this is likely to expand to include more unmanned aircraft systems (UAS), more use of AI and ML in logistics, data analysis, and simulating environments, as well as in intelligence, surveillance and reconnaissance (ISR) more broadly. If procured and deployed when they are safe, secure and derisked, these systems and their new capabilities carry the promise of less harm to members of the armed forces or civilians, cost-saving, and more information and control.

2.4 However, if these systems are prematurely procured and deployed, they could introduce several new vulnerabilities. For example, a new UAS that was prematurely procured and deployed may fail in unexpected ways, harming its operator or civilians. It may be vulnerable to certain adversarial attacks. And systemic risks may arise from the interaction of unsafe systems, leading to unintended escalation and increased uncertainty. Paul Scharre vividly describes this risk of a ‘flash war’ [10] - analogous to stock market ‘flash crashes’ caused by the interaction of AI and ML systems, which can cost millions of dollars but are unpredictable and unexplainable by humans.

2.5 The vulnerabilities introduced by premature procurement are:

- Most AI and ML systems are developed in the civilian private sector, which operates in a development environment with a relative absence of adversaries. This environment lacks the security mindset in which many defence systems have been developed, and there may be adversarial threats to which the system is unlikely to be designed to resist. [will be deployed in environments with well-resourced adversaries]
- Additionally, these systems sit on top of new computing infrastructures (software and hardware) that have not been designed with security in mind.
- These systems are new even for the private sector, and many currently known safety and security problems remain unsolved (such as distributional shift[1] or adversarial

[1] Ensuring that a ML system recognizes, and behaves robustly in, an environment different from its training environment. [21]

examples[2]). Further safety and security problems, currently unknown, will likely emerge as these systems become more capable.

- Systems that integrate increasingly capable AI and ML introduce new challenges and exacerbate existing challenges in human-machine interaction and user experience . Well known challenges such as automation bias [11], the tendency to over-rely on automated and autonomous systems, are more acute with these more sophisticated systems. Increasing complexity adds another layer between decision makers and reality.
- AI and ML based systems affect systemic stability through speed and proliferation. They operate at high speeds, which limits the ability of people to intervene. This raises the risk of miscalculation and escalation. And the software aspects are easy to proliferate, increasing the range of actors and empowering smaller and weaker actors.

2.6 One might expect that as these risks are so substantial, premature procurement would be avoided. However, **we are concerned that the current procurement system is ill-equipped to mitigate these risks.**

2.7 While some of these vulnerabilities are known within militaries, many are not. In particular, where there is a poor connection between expertise and procurement there is the possibility of deferring to the viewpoint of the private sector developers. While militaries are traditionally risk-averse, they are also afraid of missing out or falling behind rivals. And while militaries are especially risk-averse with respect to strategic operations, infrastructure-level technology like these defence systems entangles tactical and strategic domains.

2.8 An illustrative example is provided by two workshops we co-hosted in late 2019 on “Machine Learning and Strategic Stability”. They focussed on the next 10 years and were made up of former senior defence officials, politicians, and NGO security experts, as well as academics and military officers. They included a ‘procurement exercise’ in which the participants were presented with five ‘sales pitches’ for defence systems that integrated more advanced AI and ML, and were asked to develop an investment portfolio. The participants procured the defence systems despite the substantial risks (described above) that were presented. The three arguments for doing so given by the participants were international competition, military need, and dismissing or downplaying the technological risks.

2.9 Several factors can lead to premature procurement, many of which are within established defence procurement and therefore capable of being controlled and improved. One key factor is the quality of the risk assessment and the expertise of those taking the procurement decision. Another is whether good governance structures are in place, or whether there are accountability gaps.

[2] An input (perhaps imperceptibly different) to a ML system that an attacker has intentionally designed to cause the model to make a mistake. [22]

2.10 Other internal factors include the extent to which: research and development is public-sector led or private-sector led, development initiatives are defence-forward or defence-adapted, and how heavily cost-savings are prioritised. However some factors are outside the procurement infrastructure, such as how costly ensuring safety is, whether the nation is in wartime or peacetime, and whether adversaries expose their domestic procurement practices.

2.11 Given that many of these factors are within established defence procurement, there are mitigation strategies that the MoD can put in place to reduce the risk of the premature procurement and deployment of these systems.

3. Recommendations

A. Improve systemic risk assessment in defence procurement

3.1 Systemic risk assessment at the procurement stage should include the following questions:

- Was the technology in question first developed in the private sector? If so, what characteristics of the development environment could contribute to potential vulnerabilities in the system e.g. unaccounted adversarial threats? What are these potential threat vulnerabilities? Which threats has the system been designed to resist, and which threats has it not been designed to resist?
- If all systems of this type (or critical subcomponents thereof) fail at the same time, what would be the effect on national defence?
- If a ‘black-box’ system[3] or an upstream supply chain phase is compromised by an adversary, what is the worst thing they could do?
- How might information produced by the system affect assessment of the strategic situation?
- Would information provided by the system (or all systems of a similar make) be sufficient to indicate the existence of a threat that would change strategic posture?
- Would long-term interaction with the system likely lead to loss of skills or development of dependency by its operators or clients?
- What are the main threats if the system leaked information to adversaries?
- What are the main threats if the system, blueprints of the system, or the existence of the system becomes known to adversaries?
- How will the developer communicate the limits of the safe and secure operating environment to commanders and operators?
- How will the developer map, limit and communicate surprising failure modes of the system?

[3] A system whose operations are not visible to, or not easily explainable to, a user.

3.2 By including these questions in a risk assessment process, the UK government can build appropriate caution for the procurement of emerging defence technologies - allowing for the system safety to first surpass risk. Special caution should be paid to the inclusion of AI and ML based systems into NC4ISR (nuclear command, control, communications, computers, intelligence, surveillance and reconnaissance) [12].

B. Ensure clear lines of responsibility so that senior officials can be held responsible for errors caused in the procurement chain

3.3 In addition to the accountability of operators, commanders and developers, advocates for new and experimental systems and contracts should be accountable throughout the lifetime of the systems procured.

3.4 To ensure that there is clear, delineated collaborative responsibility there needs to be a shift away from a focus on mandatory human-in-the-loop operators for any autonomous defense system. This focus, which is dominant nationally and internationally, tends to narrow the scope of responsibility and accountability to end-point actions and end-point operators. But responsibility should be shared by all actors engaged in the research, development, procurement, and deployment of defence technology. The UK governance schemes should codify this concept of collaborative responsibility. Without the delineation and codification, the practice of accountability for premature and unsafe defence technology would fall on a limited set of actors downstream.

3.5 Across the entire chain of a product's lifecycle, responsibility needs to reach senior levels. The lack of clear responsibilities across a system's lifecycle opens up zones of uncertainty where it is unclear which, if any, senior officials - whether that is within military groups, procurement teams, or contracting firms - are ultimately responsible for potential failings. If senior officials advocating for and procuring new systems know that they are accountable, they will thus be incentivized to reduce systemic risks at the procurement stage[4].

3.6 Certain risks will be shared across most defence and defence-adjacent systems that integrate increasingly capable AI, ML and autonomy. That is because those safety or security risks stem directly from the software and hardware that gives those systems their capabilities. This suggests a common approach towards systems of this kind could be effective and efficient.

3.7 One way to operationalise this, for example in the area of safety, would be for the Defence Safety Authority (DSA) to be given responsibility for regulating the safety of defence systems that integrate increasingly capable AI, ML and autonomy. The DSA currently has responsibility for e.g. laser safety and explosives regulations [13][14]. This could involve adopting a new regulation in the form of a Joint Service Publication (JSP). This would require a targeted increase

[4] Ensuring that responsibility reaches senior staff is expected best practice for governance in civilian areas ranging from civil nuclear security to financial services.

in funding for the DSA, and additional hiring and training that could help judge the limitations, risks, and overall safety and security of new defence systems[5].

C. Consider how shifts in international standards for autonomous systems will affect UK standards and practices, and build flexible procurement standards accordingly.

3.8 Currently, there is substantial international agreement that human-in-the-loop or "meaningful human oversight" is a core characteristic of autonomous systems, especially those capable of lethal action. The UK holds the position that no machine will autonomously decide a course of action, such as initiating a kill shot.

3.9 Shifting international standards for autonomous weapons may influence the UK's position on mandatory human oversight. We have conducted 17 expert interviews on the governance of lethal autonomous weapons systems. Through these, top officials in the United States Department of Defense have signalled the unsustainable nature of human-in-the-loop mandates in the international field. If it is technologically possible to operate without human oversight, we will likely see the implementation of fully autonomous weapons systems.

3.10 The UK must consider how a shift in the weapons standards of other governments will impact national definitions, standards, and practices: How can the UK predict under what conditions the development, procurement, and operations standards would be put under pressure? What does this mean for safe procurement? What are the foreseeable risks? Should the UK step away from mandates such as human oversight for autonomous systems, in what ways can the procurement process supplement the insurance of well-functioning defence tools?

D. Update the MoD's definition of lethal autonomous weapons systems

3.11 Within the wide set of defence and defence-adjacent systems that integrate increasingly capable AI and ML, particular attention is rightly paid to *lethal* autonomous weapons systems. These systems raise important questions of ethics and international humanitarian law, and are the focus of arms control negotiations at the United Nations.

3.12 The MoD's definition of lethal autonomous weapons systems is quite different from that used by many other nations. It is idiosyncratic as it defines an 'autonomous' system as "capable of understanding higher-level intent and direction", "capable of deciding a course of action, from a number of alternatives, without depending on human oversight and control" and "able to take appropriate action to bring about a desired state". [15] This is a very high bar to cross - almost human-level intelligence - so high as to be almost meaningless. No system currently under

[5] Work in this area could refer to and build on civilian guidance on the management of ML systems, including Regulation for the Fourth Industrial Revolution and the Center for Data Ethics and Innovation's Guide to Using Artificial Intelligence in the Public Sector.

research or development will be capable of this very advanced capability. This is out of step with the definitions used by most other governments.

3.13 This led the House of Lords Select Committee on Artificial Intelligence report from two years ago to recommend:

> “that the UK’s definition of autonomous weapons should be realigned to be the same, or similar, as that used by the rest of the world. To produce this definition the Government should convene a panel of military and AI experts to agree a revised form of words. This should be done within eight months of the publication of this report.” [16]

3.14 In the Lords Debate on the report, Lord Browne of Ladyton suggested that this proposed panel should be expanded beyond “the military and the experts” to a “broad and ongoing UK conversation” that includes MPs and Peers. He argued that the present MoD definition is problematic because:

> “It limits the UK’s participation in the international debate, because it speaks a different language; it restricts our ability to show moral and ethical leadership; and it blocks the possibility that the current international process that is considering how to control these weapons systems will reach an agreed definition”. [17]

3.15 The panel suggestion was dismissed by the Government at the time in its Response: “The Ministry of Defence has no plans to change the definition of an autonomous system.” [18]

3.16 However, over 2019 the Government’s contributions to the lethal autonomous weapons systems negotiations at the UN Convention on Certain Conventional Weapons (CCW) have shifted slightly [19]. Specifically, over the last year the Government’s representatives have suggested:

> ‘There may be merits in a ‘code of conduct’ which would ‘provide space to allow discussions to evolve towards an outcome,’ whilst ‘reducing the risks of unchecked and unregulated research and development’; and
>
> Further efforts should be dedicated to ‘operationalising’ the CCW’s Guiding Principles ‘in order to provide a LAWS-specific set of guidelines which could be overlaid on and integrated with existing regulatory structures’. [20]

3.17 These shifts are constructive and welcome. Nevertheless, the fact that the UK’s national definition is so out of step with its international allies is still holding the UK back. As Lord Browne argued, it limits the UK’s influence on the international stage, holding the UK back from leadership.

3.18 Of particular importance to this Inquiry, this definition also creates problems for the UK's procurement, defence industry and exports. It complicates the UK's procurement of weapons systems that are in line with the principles of international humanitarian law. Furthermore, it limits the UK in its ability to consider and protect against foreseeable risks associated with these systems, and to set international standards for this emerging technology. Finally, if the national definition is out of step with the UK's allies, it will limit the export market for the UK's defence systems.

3.19 The Integrated Security, Defence and Foreign Policy Review provides an excellent opportunity to update this definition.

11 May 2020

References

[1] Alexander Babuta, Marion Oswald and Ardi Janjeva. (2020). Artificial Intelligence and UK National Security: Policy Considerations. RUSI.

[2] Forrest E. Morgan, Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima, Derek Grossman. (2020). Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World. RAND.

[3] UNIDIR. (2018). The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence. UNIDIR Resources No. 8.

[4] Miles Brundage & Shahar Avin et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv:1802.07228.

[5] Miles Brundage, Jasmine Wang, Shahar Avin, Haydn Belfield & Gretchen Krueger et al. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arXiv:2004.07213.

[6] Shahar Avin & Sonja Amadae. (2019). Autonomy and machine learning at the interface of nuclear weapons, computers and people. In *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Euro-Atlantic Perspectives*, SIPRI.

[7] Amritha Jayanti & Shahar Avin. (Working Draft). It Takes a Village: The Shared Responsibility of 'Raising' an Autonomous Weapon.

[8] HMG Government. (2017). Industry for Defence and a Prosperous Britain: Refreshing Defence Industrial Policy.

[9] Philip Dunne. (2018). Growing the Contribution of Defence to UK Prosperity: A report for the Secretary of State for Defence.

[10] Paul Scharre. (2018). *Army of None: Autonomous Weapons and the Future of War*. W. W. Norton & Company.

[11] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. (2011). Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012 Jan-Feb; 19(1): 121–127.

[12] Page O. Stoutland, Samantha Pitts-Kiefer, Ernest J. Moniz, Sam Nunn, and Des Browne. (2018). Nuclear Weapons in the New Cyber Age. Nuclear Threat Initiative.

[13] Defence Safety Authority. (2019). Military laser safety (JSP 390).

[14] Defence Safety Authority. (2019). JSP 482: MOD explosives regulations

[15] Development, Concepts and Doctrine Centre. (2017). Joint Doctrine Publication 0-30.2 Unmanned Aircraft Systems.

[16] House of Lords Select Committee on Artificial Intelligence. (2018). Report of Session 2017–19 HL Paper 100 AI in the UK: ready, willing and able? HL Paper 100.

[17] Lord Browne of Ladyton. (2018). Speech. Hansard, HL Debate, 19 November 2018, Volume 794, Col 30-33.

[18] HMG Government. (2018). Government response to House of Lords Artificial Intelligence Select Committee's report on 'AI in the UK: ready, willing and able?'

[19] Article 36. (2018). Shifting definitions - the UK and Autonomous Weapons Systems. Article 36.

[20] Richard Moyes. (2020). From "pink eyed terminators" to a cleareyed policy response? UK government policy on autonomy in weapons systems. Article 36.

[21] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Man. (2016). Concrete Problems in AI Safety. arXiv:1606.06565

[22] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. (2014). Explaining and Harnessing Adversarial Examples. arXiv:1412.6572.